

12. Exercise Sheet

Pattern Recognition and Neural Networks

The solutions to the problems indicated with (*...) may be submitted until **Wednesday, January 27st, 2010**, either in the secretariat of the Lehrstuhl für Informatik VI or at the latest before the exercise lesson on the same day.

Please submit your programs as printout *as well as* via email to <{ratajczak, wiesler}@cs.rwth-aachen.de> (as a zipped directory), stating the list of contributors.

1. Cluster Analysis

The K-Means clustering algorithm is a simple non-hierarchical clustering method, where the cluster centers are randomly initialized, and then refined using EM. In this task you will use the Netlab K-Means implementation (or optionally any other K-Means implementation) to experiment with clustering for visualization and classification purposes.

- On our ftp server `wasserstoff.informatik.rwth-aachen.de` the so called 'Old Faithful' data set is available in the directory `pub/PatRec/cluster/`. It tabulates the duration in minutes for 272 eruptions of the geyser, along with the corresponding waiting times till the next eruption. Use k-means clustering to cluster the data. How should you choose the number of cluster centers and their initialization range to suit the data set? (* 4P)
- Visualize the result in a two dimensional plot, using different colors for samples belonging to different clusters, and marking the cluster centers. What are your observations. (* 3P)
- Finally repeat the experiments but use an appropriate rescaling of the axis. What differences do you observe. (* 3P)

2. EM Optimization of a Semitied Language Model

Consider the problem of training a bigram language model $p(w|v)$, i.e. the probability distribution of words w given predecessor words v where w, v are words from a vocabulary \mathcal{V} . Assume training data is given by the word sequence w_1, w_2, w_3, \dots which is represented by counts $N(w, v)$:

$$N(w, v) = \sum_n \delta_{w, w_n} \delta_{v, w_{n-1}}.$$

Usually the vocabulary size $|\mathcal{V}|$ is too high to be able to observe all different word pairs w, v resulting in a high number of word counts $N(w, v)$ being zero. A possibility to circumvent the need to handle unobserved word pairs is the following "semitied" language model:

$$p(w|v) = \sum_{c=1}^C p(w|c) \cdot p(c|v),$$

with a hidden state c which can take C different values.

- Specify the normalization conditions for $p(w|v)$, $p(w|c)$, and $p(c|v)$. (* 3P)
- Specify the *Expectation Maximization* auxiliary function $Q(\{p(w|c), p(c|v)\}, \{\bar{p}(w|c), \bar{p}(c|v)\})$. (* 4P)

- (c) Derive the reestimation formulas for the new probabilities (parameters) $\bar{p}(w|c), \bar{p}(c|v)$ (* 5P)
given the probabilities $p(w|c), p(c|v)$ from a previous iteration and the counts $N(w, v)$.
- (d) Specify on what parameters the choice of an optimal number of hidden states C does (* 3P)
depend and explain why.